Changing The Success Probability in Computerized Adaptive Testing: A Monte-Carlo Simulation on The Open Matrices Item Bank

Hanif Akhtar

Universitas Muhammadiyah Malang, Indonesia, https://orcid.org/0000-0002-1388-7347

Abstract: For efficiency, Computerized Adaptive Test (CAT) algorithm selects items with the maximum information, typically with a 50% probability of being answered correctly. However, examinees may not be satisfied if they only correctly answer 50% of the items. Researchers discovered that changing the item selection algorithms to choose easier items (i.e., success probability > 50%), albeit not optimum from a measurement efficiency standpoint, would provide a better experience. The current study aims to investigate the impact of changing the success probability on measurement efficiency. A Monte-Carlo simulation was performed on the Open Matrices Item Bank and simulated item bank. A total of 1500 examinees were generated. We modified the item selection algorithm with the expected success probability of 60%, 70%, and 80%. Each examinee was assigned to five item selection methods: maximum-information, random, p=0.6, p=0.7, and p=0.8. The results indicated that traditional CAT was 60-70% shorter than random item selection. Altering the success probability did not affect the estimation of the examinee's ability. Increasing the probability of success in CAT increased the number of items required to achieve specified levels of precision. Practical considerations on how to maximize the trade-off between examinees' experiences and measurement efficiency are mentioned in the discussion.

Keywords: Adaptive Testing, Easier CAT, IRT, Open Matrices Item Bank, Test-Taking Experience

Citation: Akhtar, H. (2023). Changing The Success Probability in Computerized Adaptive Testing: A Monte-Carlo Simulation on The Open Matrices Item Bank. In M. Koc, O. T. Ozturk & M. L. Ciddi (Eds.), *Proceedings of ICRES 2023-- International Conference on Research in Education and Science* (pp. 2116-2125), Cappadocia, Turkiye. ISTES Organization.

Introduction

Computerized Adaptive Testing (CAT) is a type of assessment that utilizes computer technology to adjust the difficulty level of the test items according to the examinee's proficiency level. This type of testing is getting more popular due to its advantages. One of the most substantial advantages of the CAT over linear tests is that it provides more precise estimates with a shorter test than traditional fixed-item testing (FIT) (Wainer, 2000; Weiss, 2011). The administration of CAT relies on the accuracy of the examinee's answers to the previously administered items to determine the selection of the subsequent item or set of items. If the examinee answers the item correctly, the next item will be more difficult. On the contrary, if the examinee answers the item incorrectly, the next item will be easier. After responding to a test item, the person's ability estimate, theta, is

updated. This cycle will continue until the stopping rule (e.g., level of precision or number of items) has been reached.

The most widely used criterion for item selection in CAT is the Maximum Information (MI) criterion (van der Linden, 2005). This particular criterion is designed to minimize the standard error of measurement (SE) by identifying the item with the highest information function value during the last estimate. Modern CATs mostly use item response theory (IRT) to calibrate items and estimate the examinee's ability. In IRT, too easy or too difficult items provide little information about that examinee's ability. The CAT algorithm chooses the difficulty of the item to match the currently estimated ability. Consequently, high-ability examinees face more difficult items, whereas low-ability examinees face easier items. Within the Rasch or 2PL-IRT framework, the examinees will have a 50% probability of being answered correctly, regardless of their ability. This optimum CAT algorithm can reduce the test length to reach the prespecified measurement precision.

Several researchers believed that utilizing the CAT item selection algorithm would result in a challenging and optimally motivating assessment scenario for the examinees (Linacre, 2000; Wise, 2014). This is due to the fact that the algorithm ensures that the examinees are not compelled to work on items that are either too difficult or too simple, thereby preventing feelings of being over- or under- challenged. However, empirical evidence supporting this claim is unclear. A recent meta-analysis study found no overall effect of test type on motivation when comparing CAT with FIT (Akhtar et al., 2022). According to Andrich (1995), a success probability of 50% may not be sufficient for maintaining test-takers motivation in a CAT. This is because they will only be able to achieve success in roughly half of the items presented to them. This success probability is typically lower than people used to in fixed-item tests.

Several researchers modified the CAT setting in order to maximize examinees' motivation. For example, Ling et al. (2017) found that participants who completed easier CAT (i.e., CAT with success probability > 50%) showed higher engagement and lower anxiety than participants who completed traditional CAT or FIT. In their study, they modified the success probability to 70%. A similar study also found that participants in the easier CAT group showed slightly lower anxiety than in traditional CAT or FIT groups, even though the difference was not significant due to low statistical power (Revuelta et al., 2003). Häusler & Sommer (2008) discovered that alterations made by utilizing base success probabilities above 70% decreased measurement precision and, more importantly, a bias in estimating person parameters for respondents with higher abilities. However, a few easier motivator items could enhance test-taking motivation throughout the test without sacrificing testing time.

Evidence indicated that modifying the CAT algorithm to present easier items (i.e., success probability > 50%) would provide a better experience for examinees. However, modifying CAT to select easier items is not without consequences. When the difficulty level of a test is lower, additional items are necessary to achieve the designated SE. In their simulation, Bergstrom et al. (1992) found that if the specified SE is 0.5, the length of the test would be 16 items for the difficult test ($p = 0.5$), 17 items for the moderate test ($p = 0.7$), and 19 items for the facile test ($p = 0.7$). In their study, the item bank was calibrated using the Rasch model. In the Rasch model,

an item with p = 0.5 will always provide maximum information. However, in the 2PL model, p = 0.5 is necessary but not sufficient to provide maximum information. Therefore, practical considerations should be made to maximize the trade-off between test-taking experience and measurement efficiency.

Study objectives

This study aimed to investigate the impact of changing the success probability on measurement efficiency. I manipulated the success probability of administered items to 60%, 70%, and 80%. As a baseline, I also simulated traditional CAT using MI item selection and non-CAT using random item selection. A real and simulated item bank was used for this study. Specifically, our research questions were as follows:

1. Does changing the success probability detrimental to measurement efficiency (i.e., longer test length)? Specifically, at what point does measurement efficiency decrease drastically?
2. Does the estimates ability differ when different item selection methods are applied?

This study was divided into two stages. The first stage was a simulation on a simulated item bank. The same procedure was then replicated to a real item bank to examine whether item bank characteristics influenced the results. I used the Open Matrices Item Bank (OMIB, Koch et al., 2022) for the real item bank since it provides many benefits. The OMIB provides free and unlimited access to a large set of empirically validated figural matrices items. Therefore, researchers who are interested in applying easier CAT in a real testing context would not face any access barrier.

Methods

Item bank

This study used two item banks: simulated item bank and real item bank. All item banks used the 2PL model for item calibration. A simulated item bank was generated through R software (R Core Team, 2012) to generate an item bank as realistically as possible. A simulated item bank consisted of 1000 items with varied item difficulty (*b*) and item discrimination (*a*). Item difficulty for the simulated item bank was normally distributed. The mean and standard deviation of the distribution on the log scale for item discrimination were set as 0.2 and 0.3, respectively. It resulted in the final simulated item bank with *b* parameter (M = 0, SD = 1, Min = -3.39, Max = 3.66) and *a* parameter (M = 1.26, SD = 0.38, Min = 0.49, Max = 3.14). The distribution of *a* parameter and *b* parameter of simulated item bank is shown in Figure 1.

The second item bank used for the simulation is the real item bank from Open Matrices Item Bank (OMIB, Koch et al., 2022), freely available at https://osf.io/fqtzp. The bank consisted of 220 items. The final distribution of the OMIB had *b* parameter (M = -0.17, SD = 0.99, Min = -8.98, Max = 2.41) and *a* parameter (M = 2.09, SD = 0.84, Min = 0.11, Max = 5.16). Figure 1 depicts the distribution of *a* parameter and *b* parameter of OMIB.

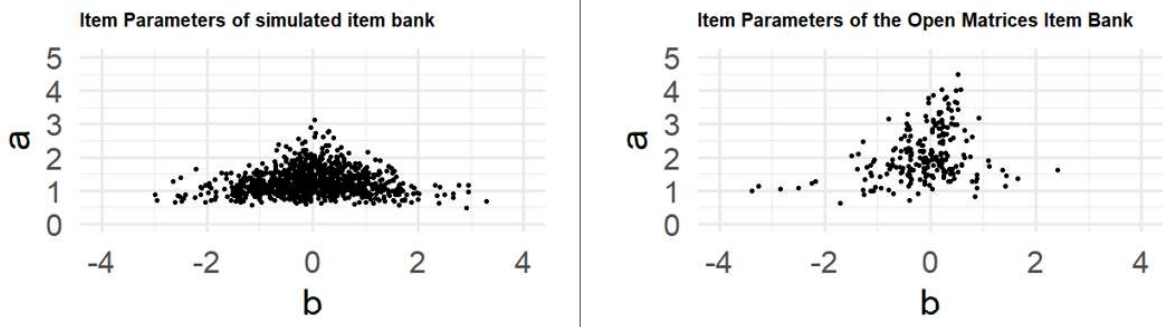

Figure 1. Distribution of item parameters of the item banks

Design of the CAT Simulation

CAT simulation was performed in two stages. In stage 1, simulation was performed in the simulated item bank. Examinee response and theta were generated using the mirtCAT package (Chalmers, 2016). The theta parameters were drawn from a standard normal distribution (M=0, SD=1), and the sample size was fixed to 1500 for each condition. The simulation design differed in terms of item selection and stopping rule. Five different item selection methods were used: Maximum-information (MI), success probability 60% (p = 0.6), success probability 70% (p = 0.7), success probability 80% (p = 0.8), and random items selection methods (Non-CAT). As a stopping rule, the precision-based termination rules were utilized with two conditions: SE < 0.30 (analogous to high-stakes testing) and SE < 0.40 (analogous to low-stakes testing). For ability estimation, Bayesian Maximum A Posteriori (MAP) was used. In stage 2, the same procedures were replicated in the OMIB.

Three criteria were used to evaluate the simulation for all conditions: test length, root means square error (RMSE), and correlation between estimated and true theta (r_{xt}). The test length was simply the number of items required to terminate. Test length was the main interest of this study as it was the indicator of measurement efficiency. RMSE was the absolute difference between the estimated and true theta. The r_{xt} is the Pearson correlation between the estimated and true theta. In addition, I also compared the real proportion of correct answers for each condition to check whether the item selection algorithm worked as expected.

Results

Measurement efficiency of each item selection

To answer the first research question (*Does changing the success probability detrimental to measurement efficiency?*), I compared all conditions in the simulation. The complete findings of the simulation study are summarized in Table 1. First, simulation was performed on the simulated dataset. As predicted, the modification from MI item selection to success-probabilities-based item selection increased the test length. For p = 0.8, the

test length was even worse than using random item selection. However, RMSE and r_{xt} in all item selection methods did not differ much. Traditional CAT (MI item selection) generally was 70% shorter than non-CAT, while moving to success-probabilities-based item selection decreased the efficiency to 13%-20%. The proportion of correct answers indicated that the real success probabilities were near to what was expected from the algorithm.

Table 1. Results of the simulation study

Item selection	SE < 0.30				SE < 0.40			
	Mean k	Prop	r_{xt}	RMSE	Mean k	Prop	r_{xt}	RMSE
Simulated item bank								
MI	11.72	0.52	0.95	0.31	6.36	0.54	0.91	0.42
p = 0.6	31.92	0.59	0.95	0.3	17.52	0.6	0.91	0.41
p = 0.7	34.86	0.68	0.95	0.3	17.96	0.68	0.91	0.41
p = 0.8	42.03	0.76	0.95	0.32	22.69	0.76	0.91	0.42
Random	40.18	0.5	0.95	0.3	20.97	0.5	0.91	0.4
Open Matrices Item Bank								
MI	11.06	0.56	0.95	0.3	5.34	0.57	0.92	0.38
p = 0.6	18.93	0.61	0.95	0.3	8.86	0.62	0.92	0.39
p = 0.7	21.08	0.67	0.95	0.3	10.42	0.68	0.91	0.4
p = 0.8	24.07	0.73	0.96	0.29	11.42	0.75	0.92	0.38
Random	26.92	0.53	0.95	0.3	13.63	0.55	0.92	0.39

Note: mean k = average test length, prop = proportion of correct, r_{xt} = correlation between the estimated and true theta, RMSE = root means square error, MI = maximum information, p = success probability.

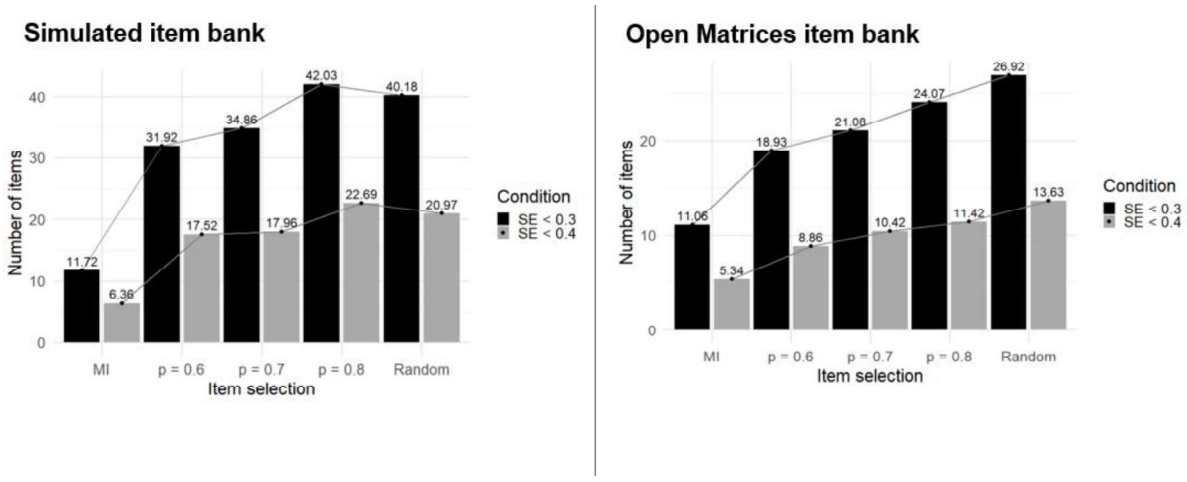

Figure 2. Number of items in each item selection method

The simulation study on OMIB showed similar results. However, modifying the algorithm using success-

probabilities-based item selection did not increase test length drastically, as in the simulated dataset. Using success-probabilities-based item selection was still better than using random item selection. Traditional CAT was generally 60% shorter than non-CAT, and its efficiency decreased to 21%-34% when success-probabilities-based item selection was used. The trend of the increasing number of items in each item selection method is displayed in Figure 2. As shown in Figure 2, when the algorithm switched from MI to p = 0.6, the test length increased drastically, but then the switching from p = 0.6 to p = 0.7 did not increase too much.

Estimate's ability in each item selection

To answer the second research question (*Does the estimates ability differ when different item selection methods are applied?*), I compared the estimated theta from each item selection method. The comparison was performed using one-way ANOVA. In any comparison, the estimated theta did not differ significantly. For the simulated item bank, in condition SE < 0.30, there was no effect of item selection on estimated theta ($F(4, 2495) = 0.17$, $p = 0.95$), and $F(4, 2495) = 0.80$, $p = 0.52$, for SE < 0.40 condition. For the OMIB, in condition SE < 0.30, there was no effect of item selection on estimated theta ($F(4, 2495) = 0.32$, $p = 0.86$), and $F(4, 2495) = 0.40$, $p = 0.81$, for SE < 0.40 condition. Figure 3 shows the box plot of thetas in each condition.

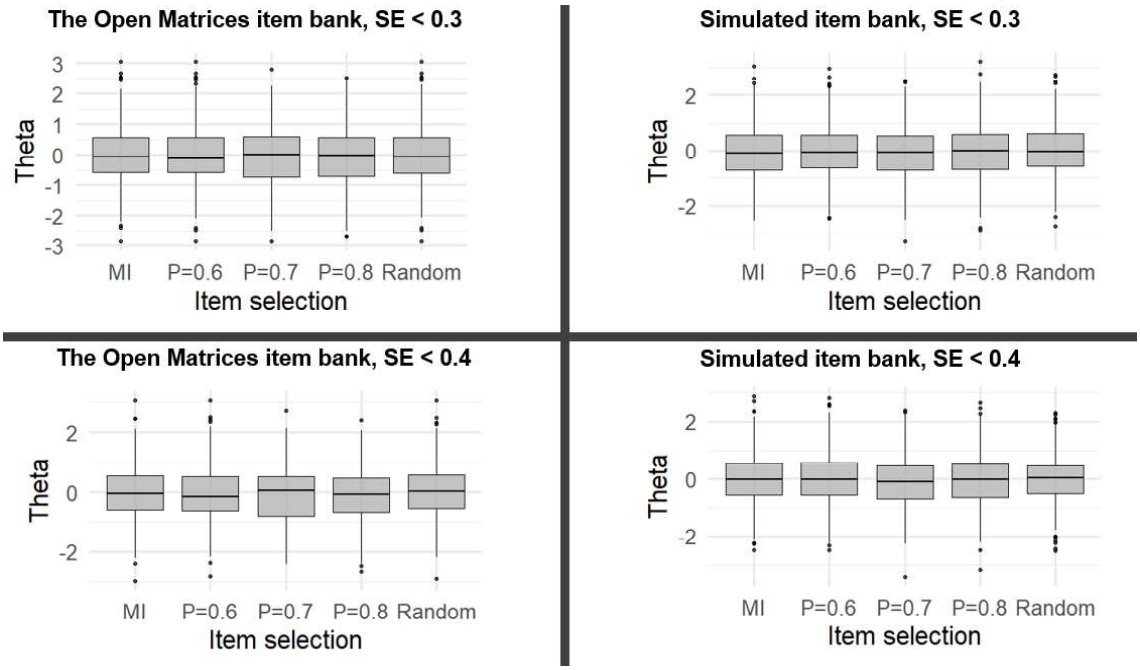

Figure 3. Comparison of estimated ability in all conditions

Discussion

This study aimed to examine the impact of changing the success probability on measurement efficiency and estimated ability. This study's main findings indicate that modifying the CAT algorithm to choose easier items

(i.e., success probability > 0.5) decreases the measurement efficiency. The test length drastically increased when p = 0.6 was applied but then was relatively stable when p = 0.7 was used. These results were similar for the simulated item bank and OMIB. However, a modifying algorithm to choose easier items did not affect r_{xt}, RMSEA, and estimated ability. Therefore, even though examinees answered more items correctly than in traditional CAT, their final theta was similar to when tested using traditional CAT.

The simulation study results have been predictable since previous studies found the same (Bergstrom et al., 1992; Häusler & Sommer, 2008). However, this study provides additional insight since two different item banks were used. First, although the OMIB consists of fewer items than the simulated item bank, it provides more efficient testing. It is particularly true when a modification algorithm is used for item selection. For example, to reach SE < 0.3, when p = 0.7 was applied, the simulated item bank needed 34.86 items, while the OMIB only needed 21.08 items. It is plausible since OMIB has more easy items (i.e., $b < 0$) with high a parameter than the simulated item bank (see Figure 1). In the IRT framework, items with higher a paramater provide maximum information, which in turn reduces SE significantly. Thus, measurement efficiency is not solely determined by the number of items in the bank, but rather by the quality of the item bank itself. However, since the simulated item bank has more items, the gap between the targeted and actual proportion of correct answers is closer than in the OMIB.

Second, modifying the algorithm to select easier items has no impact on r_{xt} and RMSEA. It indicates that, although not optimal from measurement efficiency, selecting easier items does not decrease measurement accuracy. This idea is also supported by the finding that the estimated ability (theta) did not differ in all item selections. This is true for the two different item banks. Therefore, researchers and practitioners who intend to apply easier CAT should not fear the loss of accuracy.

One notable finding of this study that contradicts previous research was the test-length cost when MI algorithm was replaced by selecting easier items. A previous study (Bergstrom et al., 1992) suggested that easier CAT targeted at a success probability of p = 0.7 only slightly increases the number of items required to reach a certain level of measurement precision. However, this study indicated that CAT targeted a success probability of p = 0.7 increases the number of items twice as CAT with MI item selection. It should be noted that Bergstrom et al. (1992) used the Rasch model. In the Rasch model, all items have similar item discrimination. Therefore, all item has the same weight to provide information as long as they match the examinees' estimated theta. In our study, the 2PL model was employed. Although theoretically, MI will result in p = 0.5, selecting items based solely on targeted p = 0.5 does not always result in maximum information. When item difficulty matches the estimated theta, item discrimination will highly determine information. Since the item bank has many items with high item discrimination (i.e., a parameter > 1), it results in more efficient testing. Thus, switching to item selection with a higher success probability will reduce efficiency even more severely.

Several scholars suggested that practitioners need to consider easier CAT because it provides better experiences for examinees (Häusler & Sommer, 2008; Ling et al., 2017; Revuelta et al., 2003). However, what is the cost if

we modify the CAT algorithm to select easier items? This study shows that the test length cost is more severe than in previous studies. Practical considerations should be made to maximize the trade-off between test-taking experience and measurement efficiency. If measurement efficiency is the priority, then no need to move from traditional CAT using MI item selection. In fact, the actual proportion of correct answers using MI was higher than 50%. But, if test-taking experiences are the priority, I would suggest item selection with targeted p = 0.70. The simulation indicated that CAT targeted at a success probability of p = 0.7 only slightly increases the number of items required to reach a certain level of measurement precision compared to p = 0.6. This suggestion is also supported by Asseburg and Frey (2013), who suggested that a success probability of 70% is optimal because examinees invested more effort and reported less boredom.

Limitations

Notably, although the central issue of this study is to address the trade-off between measurement efficiency and test-taking experience, I did not examine the effect of item selection on test-taking experiences using real-time testing. It is imperative to address whether examinees prefer shorter but harder tests or easier but longer tests. Second, I only used the 2PL model for the simulation. Other models, such as Rasch, might provide different results. For instance, in 2PL, MI item selection prefers to choose items with high *a* parameter. But in Rasch, *a* is constrained to be 1 for all items. Thus, efficiency mainly depends on the distribution of item difficulty. Third, I only used MI for traditional item selection in CAT. However, many other item selection methods exist, such as the Kullback-Leibler criteria. Future research is suggested to compare other methods too. Finally, measurement efficiency in this study was defined as the test length. However, the longer test does not always result in a longer testing time. Hornke (1997, 2000) indicated that incorrect items require longer than correct ones. Thus, easier CAT, although it needs more items, does not necessarily require more testing time. Future studies should investigate further on this issue.

Conclusion

In summary, changing the CAT algorithm to select easier items (i.e., success probability > 0.5) reduces measurement efficiency. These findings were comparable for the simulated item bank and OMIB. However, a modifying algorithm to choose easier items did not impact r_{xt}, RMSEA, and estimated ability. As a result, even though examinees accurately answered more questions than in a traditional CAT test, their final theta is comparable.

Recommendations

Practical considerations should be made to optimize the trade-off between the test-taking experience and measurement efficacy. There is no need to switch from the traditional CAT using MI item selection if measurement efficiency is the top priority. In reality, more than 50% of the questions were answered correctly

using MI. However, if the quality of the test-taking experience is more important, I would advise choosing items with an intended p = 0.70.

References

Akhtar, H., Silfiasari, Vekety, B., & Kovacs, K. (2022). The Effect of Computerized Adaptive Testing on Motivation and Anxiety: A Systematic Review and Meta-Analysis. *Assessment*, 10731911221100996. https://doi.org/10.1177/10731911221100995

Andrich. (1995). Review of the book Computerized Adaptive Testing: A Primer. *Psychometrika*, *4*, 615–620.

Asseburg, R., & Frey, A. (2013). Too hard , too easy , or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, *55*(1), 92–104.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the Level of Difficulty in Computer Adaptive Testing. *Applied Measurement in Education*, *5*(2), 137–149. https://doi.org/10.1207/S15324818AME0502_4

Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software*, *71*(5), 1–38. http://dx.doi.org/10.18637/jss.v071.i05

Häusler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science Quarterly*, *50*(1), 75–87.

Hornke, L. F. (1997). Investigating item response times in computerized adaptive testing. *Diagnostica*, *43*(1), 27–39.

Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicológica*, *21*, 175–189.

Koch, M., Spinath, F. M., Greiff, S., & Becker, N. (2022). Development and Validation of the Open Matrices Item Bank. *Journal of Intelligence*, *10*(3), 41. https://doi.org/10.3390/jintelligence10030041

Linacre, J. M. (2000). Computer-Adaptive Testing: A Methodology Whose Time Has Come. By J ohn Michael Linacr e , Ph. D. MESA Psychometr ic Labor ator y Univer sity of Chicago. *Test*, *69*, 27.

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

R Core Team. (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Revuelta, J., Ximénez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement*, *63*(5), 791–808. https://doi.org/10.1177/0013164403251282

van der Linden, W. J. (2005). Item Response Theory. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (pp. 379–387). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00452-7

Wainer, H. (2000). *Computerized adaptive testing: A primer (Second edition)* (2nd ed.). Lawrence Erlbaum Associates.

Weiss, D. J. (2011). Better Data From Better Measurements Using Computerized Adaptive Testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1. https://doi.org/10.2458/jmm.v2i1.12351

Wise, S. L. (2014). The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees. *Journal of Computerized Adaptive Testing*, *2*(1). https://doi.org/10.7333/1401-0201001